OVERVIEW OF THE ARISE PROJECT

Els den Os[1], *Lou Boves*[1,2], *Lori Lamel*[3], *Paolo Baggia*[4]
[1]KPN Research, P.O.Box 421 2260 AK Leidschendam, The Netherlands
[2]University of Nijmegen, P.O.Box 9103, 6500 HD Nijmegen, The Netherlands
[3]LIMSI-CNRS, BP 133, 91304 Orsay, France
[4]CSELT, Via Reiss Romoli 274, 10148 Torino, Italy
e.a.denos@research.kpn.com, boves@lands.let.kun.nl, lamel@limsi.fr, Paolo.Baggia@cselt.it

ABSTRACT

The LE-3 project ARISE (Automatic Railway Information Systems for Europe) ran from October 1996 through December 1998. Four prototypes of train timetable information systems for three different languages were developed, tested, and validated: Italian (based on technology developed by CSELT), French (two systems, one based on Philips technology, the other on technology by LIMSI), and Dutch (based on Philips technology). The goal of ARISE was to improve the basic technology and to enhance our general understanding of the issues involved in actual deployment of spoken language dialogue systems for restricted domain information. This paper summarises the main findings of the project in terms of speech recognition, dialogue state dependent language modelling, dialogue control, information presentation, system output, evaluation of spoken dialogue systems, and operational issues.

I. INTRODUCTION

Since the comparison of a number of prototypes of spoken dialogue systems during Eurospeech-'97 [1] the number of laboratory and operational spoken language information system has steadily grown. Yet, our understanding of many of the basic issues in developing, optimising and deploying these systems is still incomplete.

In the ARISE project we have attempted to improve this situation by developing and evaluating six systems in parallel, all in the domain of train timetable information. Three systems were developed in France, two in Italy and two in the Netherlands. One of the French systems (developed by IRIT) and both Dutch systems are based on technology originally developed by Philips [2]. During the project, one Dutch system (VIOS) became operational, the other remained a research system. The second French system was developed by LIMSI[1] and Vecsys [3]. There were two Italian systems, both developed by CSELT, one isolated word recognition (IWR) system and the other a continuous speech recognition (CSR) system [4].

The research in ARISE was focused on issues in dialogue management, integration of spoken dialogue systems in an operational service, and evaluation and validation of these systems. In addition, work has been done on the improvement of the basic speech recognition and language understanding technology that is needed to build spoken dialogue systems (confidence measures, state dependent modelling, and barge-in). Not all partners explored the same problems, but by working together the project as a whole was able to compare approaches when multiple sites addressed the same problem and to address more aspects than an individual site could do alone. The collaboration may be considered highly productive, if only because the close comparison of the findings has allowed us to come to conclusions that we think are generally applicable in the domain of timetable information, and probably also in other spoken dialogue information systems.

[1] LIMSI also developed a bilingual French/English system for high speed trains between Paris and London; this is not covered in this paper.

II. FUNCTIONALITIES

Two types of functionality can be distinguished in spoken language dialogue systems, viz., the coverage of the domain and the dialogue behaviour.

The LIMSI system has the most elaborate domain coverage: it provides timetable information between 600 stations, and in addition it gives information on fares, including reductions, and on-board services. It also allows the caller to make simulated reservations on the trains. The IRIT system only gives timetable information for intercity connections between 300 stations. The Italian IWR system provides timetable, fare and on-board services information for connections with less than five changes between 664 major stations. The Italian CSR system provides the same information but for all 3,000+ stations in the Italian rail network. The Dutch system provides timetable information for the complete Dutch rail network (about 500 stations). Reservations are not possible on domestic trains, and on-board services are not scheduled.

Services aimed at the general public, like timetable information systems, cannot count on occasional users intuitively knowing the details of the system's functionality. Eventually, frequent users will get accustomed to the peculiarities in the system's dialogue behaviour. It is not possible to explain in detail the system's behaviour to the users, either. In the Dutch VIOS system callers were originally offered the option to listen to a short explanation of how to interact with the system. It appeared that callers who listened to this information were no more successful than listeners who choose not to listen to it. An experiment was run with the French IRIT system in which three different initial prompts were compared. These initial prompts were meant to help people to use the system (e.g. do not hesitate to correct the system in case of errors). No benefits of this information were observed in the tests. For this type of domain, information on how to interact with the system does not seem to help the callers.

Therefore, the two operational systems (in Italy and the Netherlands) do not provide help information at the dialogue start. However, rephrasing of the initial prompt may have effects on the behaviour of the callers. In the LIMSI system changing the initial prompt from "how can I help you?" to "what information do you want?" resulted in more informative first queries (cf. also [2]).

III. SPEECH RECOGNITION

Analysis of successful and failed calls clearly shows that most problems are caused by persistent recognition errors. An additional problem is that users often do not understand that miscommunication is the result of speech recognition errors. Consequently, it is our impression that the lion's share of dialogue management intelligence in the present generation of systems is mainly needed to cope with recognition errors. Below, we present some of the solutions for recognition errors that we tried in the ARISE systems.

For the continuous speech systems most of the fatal recognition errors were related to station names. Problems with station names will persist for a long time to come, if only because of the presence of names that are easily confused, even by human listeners. In the LIMSI system station names can be optionally spelled. If no city is detected in the user's query, it prompts first with an example (with no change of the recognition models). If there is still no response, the prompt "give your arrival/departure city and spell it if you want, for example Paris, P A R I S" is played. The reply is processed with more precise acoustic models and a constrained language model. Combined with rejection of uncertain words this has led to improved dialogue success rate.

All but the Italian IWR system use continuous speech recognition. The Italian IWR system, with a city name vocabulary of 664 words, reached 92.7% correct name recognition in a test with real users. The Italian CSR system switches to isolated word recognition in case of persistent negations.

The French system built by IRIT and the Italian CSR system keep track of dialogue history, so that they do not make the same confusion twice in a row. In the continuous speech systems the number of fatal problems with date and time expressions was smaller than in the Italian IWR, where date recognition suffered from a relatively high 17.8% out-of-vocabulary expressions, mainly because the callers seem to have misinterpreted the exact meaning of the question asked by the machine. Apparently, there is a trade-off between in-grammar recognition accuracy and the proportion of in-grammar utterances.

IV. DIALOGUE STRATEGIES

In this section we summarise the most important experiments and findings in the realm of dialogue strategies.

System-driven

By necessity, the Italian IWR system uses a completely system-driven dialogue: the system asks directed questions, which the user must answer. In this strategy the order in which the slots in the query form are filled is always the same, and determined by the system. In the timetable domain callers do not seem to have problems in complying with the system driven interaction style.

Mixed initiative

All other systems implement a mixed initiative dialogue strategy, in which the caller may volunteer to provide any information relevant to the application at any time. The LIMSI system differs from the four other systems in that its opening question is really open: *"What information do you want?"*, whereas the four other systems use a much more directed prompt: *"From where to where do you want to travel?"*. The LIMSI system cannot use this directed opening prompt, because it handles price information and reservations in addition to schedule information.

Explicit prompting

When recognition errors are detected, all systems except the Dutch VIOS system switch to more directed prompts. The two French systems and the Italian CSR system also adapt the recogniser language models, to better reflect the expected answers to the directed questions; the Dutch research system leaves its vocabulary and language model unchanged. From the data that we have available, we cannot conclude that vocabulary and language model adaptation did indeed improve system performance. Since we are aiming to improve the overall dialogue success rate, changes in acoustic and language models are often combined with other changes in the dialogue, which makes it difficult to tease out the effects. Off-line tests show that model adaptation improves performance [5], but these tests do not necessarily reflect the behaviour of the total system in a real interactive dialogue.

Confirmation strategies

One of the basic operations in any viable dialogue model is confirmation. In human-human dialogues the participants constantly monitor their mutual understanding. For instance, an operator may have understood that the caller is looking for a connection on the following day; then, if a following utterance seems to imply that the caller wants to travel on the same weekday, but one week later, a data conflict arises that must be resolved. This can be done by means of an explicit question or by asking the caller to respecify the conflicting information. The LIMSI system attempts to detect and resolve such conflicts using the latter approach. The other systems avoid conflicting data by closing slots in the query form as soon as the contents has been confirmed by the caller, implicitly or explicitly. This was done to avoid loops in the dialogue, which might occur because of recognition errors later on.

Because the vulnerability of the speech recogniser, per-utterance confirmation of new information items is virtually inevitable. This can be done in several different ways. In the Dutch VIOS system and in the French IRIT system implicit confirmation is used by default. However, it soon appeared that callers had difficulty in grasping the system's strategy that requires immediate correction of recognition errors, because the failure to make a correction is interpreted by the system as a confirmation.

One way to overcome this mismatch between the callers' perception of the system capabilities and the actual system strategy is to revert to explicit confirmation. Although this almost doubles the number of turns, it was found that explicit confirmation does not necessarily lengthen the time to task completion, if only because explicit confirmation is easily and naturally combined with short prompts [6]. Still one can do better, provided that the recogniser computes confidence measures for the words in its output. In the latest versions of the LIMSI and the Dutch research systems confidence measures were used. In the Dutch system these prove to shorten the dialogue by using implicit confirmation if the risk that an error is committed is very low [7].

Wording of implicit confirmation

There are several ways in which implicit confirmation requests can be formulated, i.e. in complete sentences like "At what day do you want to travel from A to B?", and an alternative that is used in the LIMSI, the Italian CSR system and the latest Dutch research system:

S: *From where to where do you want to travel?*
C: *From Paris to Bordeaux.*
S: *From Paris to Bordeaux, when do you want to go?*

Expert evaluation of the latter formulation in the Dutch system has shown that the prosody is critical. If the intonation in the confirmation part of the prompt is too final, callers get the impression that the system is not willing to repair errors. In any case, this formulation leads to shorter prompts, which may be an advantage on its own.

IRIT has implemented 'semi-implicit' confirmation, using formulations like *"You want to leave from Toulouse. If this is wrong, please correct. Else, say your destination."* Experiments have shown that this formulation avoids some of the misunderstandings with truly implicit confirmation, but they tend to lead to lengthy dialogues.

V. SYSTEM OUTPUT

It has been observed that the output of a spoken dialogue system is at least as important in determining the user's appreciation as speech recognition performance or the system's functionality [1]. Output generation comprises two somewhat independent processes. First, the precise wording of the system's output must be decided; second, the resulting expression must be converted into sound.

Prompt formulation

None of the systems used advanced language technology to optimise the formulation of the system prompts. In a small domain like train travel information, all relevant messages can be predicted in advance. Thus, even if one wants to allow for context dependent variation of the output, straightforward pattern combination is adequate.

In the Dutch research system an effort has been made to shorten all system prompts as much as possible. This was done in part to compensate for the higher number of turns, due to explicit confirmation. However, subjects appreciated the short utterances in their own right. In the LIMSI system concise prompts were also used, providing only new or highly relevant information.

It has been suggested that the systems should adapt to the formulations the callers use to phrase their queries, especially for time expressions. However, operators tend to use the less ambiguous 24-hour clock throughout. Thus, complicated control structures in the output generation might prove not to be cost-effective.

Speech synthesis

Experiments in ARISE have shown that even the best general purpose text-to-speech (TTS) synthesis available today is not good enough for use in a timetable information service for the general public. With the Dutch research system (which uses concatenation of phrases generated by a TTS system) as the only exception, all systems use some form of concatenation of pre-recorded natural speech. In the LIMSI system, and in both Italian systems, considerable effort was spent in recording multiple tokens of all words and phrases to make them compatible with the prosodic context in which they had to eventually be placed. This resulted in excellent quality synthetic speech. The IRIT system paid less attention to prosodic optimisation; consequently, the output quality was considered at best as acceptable.

In the near future the operational Italian system will combine concatenation of natural speech for the query part of the dialogue and TTS, based on sub-word unit concatenation optimised for this specific application, for the presentation of the travel advice.

VI. NEGOTIATION

All systems start out presenting the 'best' connection, given the query data and a set of optimisation rules. In the Dutch research system quite complicated weighting is used to trade distance to the specified departure/arrival time against the number of stops and changes. This is necessary because many cities have very frequent, but not necessarily equally comfortable connections. Moreover, from an analysis of the dialogues in the operator-based service it appeared that a substantial proportion of the callers may not be interested in a specific connection, but rather in the pattern of connections during a part of the day. Deriving this pattern from the schedule database can be very difficult.

Much effort was spent to enable flexible navigation and negotiation facilities in the Dutch research system. However, it is simply impossible to emulate an intelligent human operator; at the same time, it is extremely difficult to communicate the exact limitations in the system's functionality to the callers. The best solution seems to explicitly tell what the options are (earlier, later, fewer changes, information about platforms and train directions) if a user does not seem to be satisfied with the first travel advice.

LIMSI encountered similar problems. Although this system can handle reservations, this functionality is limited to a single seat per dialogue. Many callers had difficulty in accepting this unexpected limitation. On the other hand, the very open style at the start of a dialogue seemed to make it easier for the callers of the LIMSI system to ask for an earlier or later connection.

VII. INTEGRATION ISSUES

Spoken dialogue systems can be deployed to replace operators, but also to facilitate the task of the operators. The architecture of the Italian systems has been designed to allow both operator support and operator replacement. To that end two 'agents' are distinguished, one which collects the query data and a second who presents the travel advice. The second 'agent' can be used to relieve a human operator from the task of reading the travel advice.

All systems can operate in a fully automatic mode. In this mode they actually replace an operator. In fully automatic mode operator fallback can be implemented; both operational systems do this, during the normal opening hours of the service; during the night operator fallback is disabled. Operator fallback can be initiated by the caller (by pressing a DTMF key) or by the system. The latter happens if the system detects persistent failure to acquire an information item. In the research systems operator fallback is not offered.

The choice between operator support or operator replacement has far reaching implications for network and system integration. It remains to be investigated whether the operators' task can be facilitated when partial, and potentially erroneous, information acquired by an automatic system is put into the query form in the case of operator fallback.

VIII. EVALUATION

Evaluation in ARISE was focused on the performance of the systems in their interaction with callers. Different experimental designs were combined with different types of performance measures. Experiments with the research systems typically involved subjects carrying out predesigned scenarios. In addition, subjects were given the opportunity to ask queries of their own choice. The operational systems were assessed by monitoring their interactions with real paying customers.

Several types of measures have been acquired, including 'Service Success Rates' (SSR), time to task completion, and responses to questionnaires and Likert scales. In general terms it can be concluded that no single experimental set-up and no single measure is suitable for all research questions. For operational systems SSR and customer satisfaction are by far the most important measures. It is questionable, however, whether reliable satisfaction measures can be obtained with questionnaires. Repeated use of a service is probably a much better measure.

The automatic system in Italy showed an increase of the number of calls from 600,000 in October 1998 to 1,400,000 in December. In October over 60% of these calls were handled successfully. The performance has continually improved since the introduction, most likely because frequent users are getting familiar with the system. In March 1999 SSR was over 80%; this means that over 60% of the total number of calls served by the railway call centers in Italy are automated. The Dutch VIOS system handles a stable 90,000+ calls per month. Over 65% of the calls are completed successfully; most of the other calls are transferred to an operator. During the night, when operator fallback is not available, the SSR is significantly higher.

The last assessment of the LIMSI system, carried out by SNCF in November 1998, with subjects recruited via a polling company, yielded an SSR of 78.5%. If calls are eliminated in which the user made no attempt to correct the system when it made errors, SSR increases to 84.7%.

In experiments with the research systems we have found that often subjects do not care to correct recognition errors. This is probably due to the fact that they do not really need the information that they are requested to collect by carrying out a scenario. Moreover, we have found that it is virtually impossible to design scenarios which invite subjects to explore all functionalities of the systems without inducing the use of the same syntax and expressions as used in the description of the tasks. Graphical representations of the instructions are not suitable for more complex tasks.

IX. CONCLUDING REMARK

In the ARISE project we have learned a lot about (aspects of) Spoken Dialogue Systems in the domain of timetable information. We do not know whether the findings are applicable for other types of domains as well. Therefore, it is necessary to build systems for other domains, and investigate whether the experiences from the timetable information domain can be used.

REFERENCES

[1] Den Os, E.A. & Bloothooft, G. (1998) Evaluating various spoken dialogue systems with a single questionnaire: Analysis of the ELSNET Olympics, *Proc. First Int. Conference on Language Resources & Evaluation,* pp. 51-54.

[2] Aust, H. Oerder, M. Seide, F. & Steinbiss, V. (1994) "Experience with the Philips Automatic Train Timetable System", *Proc. IVTTA-'94*, pp. 67-72.

[3] Lamel, L., Rossett, S., Gauvain, J.J., Bennacef, S., Garnier-Rizet, M. & Prouts, B. (1998) "The LIMSI ARISE System", *Proc. IVTTA-'98*, pp. 209-214.

[4] Castagnieri, G., Baggia, P. & Danieli, M. (1998) "Field Trials of the Italian ARISE Train Timetable System", *Proc. IVTTA-'98*, pp. 97-102.

[5] Baggia, P., Gauvain, J., Kellner, A., Perennou, G, Popovici, C., Sturm, J. & Wessel, F. (1999) "Language Modelling and Spoken Dialogue Systems – the ARISE experience", *Proc. Eurospeech-'99.*

[6] Sanderman, A.A., Sturm, J.A., Den Os, E.A., Boves, L., Cremers, A (1998) "Evaluation of the Dutch Train Timetable Information system developed in the ARISE project" *Proc. IVTTA-'98*, pp. 91-96.

[7] Sturm, J., Den Os, E.A., Boves, L. (1999) "Dialogue Management in the Dutch ARISE Train Timetable Information system" *Proc. Eurospeech'99.*